QUESTION:Use Pandas to clean and preprocess a messy dataset,documenting the steps taken during the cleaning process.

PANDAS

```
#Importing the Pandas library
import pandas as pd
```

Creating a pandas dataframe from a dictionary and performing some basic operation.

```
#Creating a dictionary containing data
Data={'Name':['Maryam','Kamal','Khadija','Umar','Ibrahim','Amina'],
'Age':[23,30,18,25,35,22],
'Gender':['female','male','female','male','male','female'],
'Marital status':['Married','Married','Single','Single','Single','Single']}
```

Creating a dataframe from the dictionary

```
#Create a dataframe from the dictionary
df=pd.DataFrame(Data)
```

Displaying the dataFrame

```
#Display the dataFrame
print(df)
```

```
           Name  Age  Gender Marital status
     0   Maryam   23  female        Married
     1    Kamal   30    male        Married
     2  Khadija   18  female         Single
     3     Umar   25    male         Single
     4  Ibrahim   35    male         Single
     5    Amina   22  female         Single
```

```
#Getting information about the dataframe
print(df.info())
```

```
     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 6 entries, 0 to 5
     Data columns (total 4 columns):
      #   Column          Non-Null Count  Dtype
     ---  ------          --------------  -----
      0   Name            6 non-null      object
      1   Age             6 non-null      int64
      2   Gender          6 non-null      object
      3   Marital status  6 non-null      object
```

```
dtypes: int64(1), object(3)
memory usage: 320.0+ bytes
None
```

Filter row base on a condition

```
#Filtering column of Age greater than 20
filtered_df = df[df['Age']>20]


print(filtered_df)
```

```
           Name  Age  Gender Marital status
0   Maryam   23  female        Married
1    Kamal   30    male        Married
3     Umar   25    male         Single
4  Ibrahim   35    male         Single
5    Amina   22  female         Single
```

Filtering Marital Status from the dataframe

```
#Filtering the Single from the dataframe
filtered_df =df[df['Marital status']=='Single']


print(filtered_df)
```

```
           Name  Age  Gender Marital status
2  Khadija   18  female         Single
3     Umar   25    male         Single
4  Ibrahim   35    male         Single
5    Amina   22  female         Single
```

```
#Filtering the Married from the dataframe
filtered_df =df[df['Marital status']=='Married']


print(filtered_df)
```

```
          Name  Age  Gender Marital status
0  Maryam   23  female        Married
1   Kamal   30    male        Married
```

CLEANING A MESSY DATA USING PANDAS

Cleaning empty cells

```
#Remove Rows
import pandas as pd
```

```
df=pd.read_csv('/content/dataframe 3MTT.csv')
```

```
print(df)
```

```
                 NAMES   AGE REPORTING DATE
0  Maryam Hussain   23.0    12/02/2024'
1     Amina Kamilu  24.0            NaN
2    Bilyamin Nura  30.0    05/01/2024'
3       Kamal Musa   NaN    09/03/2024'
4        Musa Musa  25.0            NaN
5    Aisha Taufeeq  40.0    23/03/2024'
```

Removing rows that contains empty cells

```
#row 1,5,6 deleted
df.dropna()
```

	NAMES	AGE	REPORTING DATE
0	Maryam Hussain	23.0	12/02/2024'
2	Bilyamin Nura	30.0	05/01/2024'
5	Aisha Taufeeq	40.0	23/03/2024'

CLEANING DATA OF WRONG FORMAT

```
#Non-date format column
df
```

	NAMES	AGE	REPORTING DATE
0	Maryam Hussain	23.0	12/02/2024'
1	Amina Kamilu	24.0	NaN
2	Bilyamin Nura	30.0	05/01/2024'
3	Kamal Musa	NaN	09/03/2024'
4	Musa Musa	25.0	NaN
5	Aisha Taufeeq	40.0	23/03/2024'

```
df['REPORTING DATE']
```

```
0    12/02/2024'
1            NaN
2    05/01/2024'
3    09/03/2024'
```

```
4            NaN
5    23/03/2024'
Name: REPORTING DATE, dtype: object
```

CONVERTING TO CORRECT FORMAT

```
#Converting to_datatime()
df['REPORTING DATE'] =pd.to_datetime(df['REPORTING DATE'])
```

```
#Column in date format
#NaT i.e.empty cell
df['REPORTING DATE']
```

```
0   2024-12-02
1          NaT
2   2024-05-01
3   2024-09-03
4          NaT
5   2024-03-23
Name: REPORTING DATE, dtype: datetime64[ns]
```

ADDING NEW COLUMN IN DATAFRAME

```
#df[New column name]= Value
df['COUNTRY']= 'Nigeria'
```

```
df
```

	NAMES	AGE	REPORTING DATE	Country	COUNTRY
0	Maryam Hussain	23.0	2024-12-02	Nigeria	Nigeria
1	Amina Kamilu	24.0	NaT	Nigeria	Nigeria
2	Bilyamin Nura	30.0	2024-05-01	Nigeria	Nigeria
3	Kamal Musa	NaN	2024-09-03	Nigeria	Nigeria
4	Musa Musa	25.0	NaT	Nigeria	Nigeria
5	Aisha Taufeeq	40.0	2024-03-23	Nigeria	Nigeria

Next steps: Generate code with df View recommended plots

COLUMN DELETION IN DATAFRAME

```
#Using drop()
df.drop(['Country'],axis=1, inplace = True)
```

```
df
```

	NAMES	AGE	REPORTING DATE
0	Maryam Hussain	23.0	2024-12-02
1	Amina Kamilu	24.0	NaT
2	Bilyamin Nura	30.0	2024-05-01
3	Kamal Musa	NaN	2024-09-03
4	Musa Musa	25.0	NaT
5	Aisha Taufeeq	40.0	2024-03-23

Next steps: Generate code with df View recommended plots